Expert Systems With Applications 172 (2021) 114622

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Ensemble clustering using extended fuzzy k-means for cancer data analysis

Imran Khan [a,*], Zongwei Luo [c,*], Abdul Khalique Shaikh [b], Rachid Hedjam [a]

[a] Department of Computer Science, College of Science, Sultan Qaboos University, P.O. Box 31, Al-Khoud 123, Muscat, Oman
[b] Department of Information Systems at Sultan Qaboos University, P.O. Box 31, Al-Khoud 123, Muscat, Oman
[c] BNU-UIC Institute of Artificial Intelligence and Future Networks, Beijing Normal University (BNU Zhuhai), BNU-HKBU United International College Tangjiawan, Rd. JinTong 2000#, Zhuhai, Guangdong, PR China

ARTICLE INFO

Keywords:
Fuzzy *k*-means
Cluster analysis
Cancer data
Variable weights

ABSTRACT

Clustering analysis is a significant research topic in discovering cancer using different profiles of gene expression, which is very important to successfully diagnose and treat the cancer decease. Many ensemble clustering methods have been developed to perform clustering using tumor data. Only few of them incorporates a significant number of input clusterings, the optimal number of clusters in each input clustering, and an appropriate ensemble method to combine input clusterings into a final clustering. In this paper, we introduce two new steps in the standard fuzzy *k*-means algorithm to determine the optimal number of input clusterings, and the optimal number of clusters in each clustering for ensemble clustering. The first one is to incorporate a penalty term for making the algorithm insensitive to the initialization of cluster centroids. The second one is to automate a clustering process for iteratively updating the feature weights. This step addresses the noise values in the dataset. We propose an ensemble clustering method, which combines a set of input clusterings into a final clustering having better overall quality. Experiments on real cancer gene expression profiles illustrate that the proposed algorithm outperformed the well-known clustering algorithms.

1. Introduction

A most well-known problem is how to organize collected data into meaningful structures (or taxonomies) (Khan et al., 2019). Cluster analysis aims the finding of groups of instances. This is possible by finding partitions or a sequences of partitions that optimize an objective function of interest. Cluster analysis is very hot topic in many areas including medicine, biology and related areas (Handl et al., 2005; Jaskowiak et al., 2013; Maji and Paul, 2012; Panwong et al., 2020; Tautenhain and Nascimento, 2020; Yang et al., 2014). Recently, the problem of clustering bio-molecular data gets more and more attention due to its practical applications in diagnosis of cancer decease and related areas (Yu et al., 2014). The clustering process of bio-molecular data faces two major problems. One of them is to correctly determine the number of clusters from the given bio-molecular data. The other one is to successfully dividing the samples into many groups and correctly allocating the data objects to their corresponding groups.

Many cluster validity indices have been developed for analyzing the cancer data to estimation the accurate number of clusters. They are categorized into internal indices, information theory-based criteria, and external indices. Information theory-based criteria utilize entropy or likelihood measures to get the detailed structure of cancer data, e.g., the Akaike information criterion (AIC) (Shannon, 1951), the Bayesian Information Criterion (BIC) (Schwarz et al., 1978), etc. Internal indices use statistical measures to determine the number of clusters, e.g., the Silhouette Index (SI) (Kaufman and Rousseeuw, 2009), the Davies-Bouldin index (DBI) (Davies and Bouldin, 1979), etc. External indices evaluate the partitions based on stability of an individual cluster and the agreement between two clusters to estimate the number of clusters. These indices include the Jaccard coefficient (JI) (Sarle, 1990), the Rand index (RI) (Rand, 1971), etc.

Many tumour clustering algorithms have been proposed that successfully divide the samples into many groups and accurately assigning the samples to their corresponding groups. They are self-organizing map (SOM) (Yu et al., 2012), hierarchical clustering (HC) (Alizadeh et al., 2000), model based clustering (Yeung et al., 2001), nonnegative matrix factorization (Zheng et al., 2009), etc. These are successfully used to analyze the cancer data, but most of them are not stable, robust and accurate (Yu et al., 2014). The performance of the these well-known clustering algorithms is always less than that of the consensus clustering approaches, which is also referred as cluster ensemble. The ensemble clustering methods always produce more stable, robust and

* corresponding author.
E-mail addresses: i.khan@squ.edu.om (I. Khan), lzwqhk@bnu.edu.cn (Z. Luo), shaikh@squ.edu.om (A.K. Shaikh), rachid.hedjam@squ.edu.om (R. Hedjam).

https://doi.org/10.1016/j.eswa.2021.114622
Received 16 September 2020; Received in revised form 1 January 2021; Accepted 13 January 2021
Available online 26 January 2021
0957-4174/© 2021 Elsevier Ltd. All rights reserved.

accurate results (Khan et al., 2014; Khan et al., 2016; Khan et al., 2018; Monti et al., 2003; Yu et al., 2007; Yu and Wong, 2009). For instance, Yu et al. (2007) proposed a graph-based consensus clustering (GCC) algorithm perform cluster analysis on cancer data. Additionally, they proposed a new perturbation based ensemble clustering technique to assign samples to their corresponding cancer types (Yu and Wong, 2009). The two re-sampling based ensemble clustering methods are proposed by Monti et al. (2003). They used corresponding the hierarchical clustering algorithm and SOM clustering algorithm cluster the cancer gene expression profiles. Moreover, a link-based ensemble clustering technique has been developed to perform cluster analysis on cancer data (Iam-On et al., 2010).

The process of ensemble clustering comprises two steps, generating a set of individual input clusterings from a given dataset and combing the input clusterings into a final clustering (Jing et al., 2015). The robustness of the final clustering is estimated by the techniques to complete these two steps. Multiple input clusterings are generated from a dataset by.

- using various clustering methods (Strehl and Ghosh, 2002),
- using a random initialization clustering method many times (Fred and Jain, 2005),
- using a clustering algorithm multiple times with different values of input parameters (Iam-On et al., 2010), and
- using different subsets of the features in a dataset (Khan et al., 2014).

Ensemble clustering methods integrates multiple input clusterings to produce a final clustering. The obtained final clustering should be better than the individual input clusterings. The most commonly used ensemble clustering methods are Cluster-based Similarity Partitioning Algorithm (CSPA), HyperGraph Partitioning Algorithm (HPGA), and Meta-CLustering Algorithm (MCLA) (Khan et al., 2016).

Recent developments depict that clustering ensemble produces better results on gene expression datasets (Iam-On et al., 2010; Kim et al., 2009). However, the gene expression datasets have challenging characteristics. They are high-dimensional and noisy which can easily degrade the performance of clustering methods (Mimaroglu and Aksehirli, 2011). In high-dimensional data, clusters often exist in different subspaces. Ensemble clustering based on full feature space clustering algorithms fail to cluster such data (Khan et al., 2014). However, the dataset is often divided into different subsets of the available features, where each subset is used to generate an input clustering for an ensemble clustering method. The most well-known methods that generate feature subsets from of a given dataset are Random Sampling (RM), Random Projection (RP), and Stratified Sampling (SS) (Jing et al., 2015). They cannot detect the optimal number of feature subsets which causes the lack of preservation of the clustering structure in each feature subset. Finding the optimal number of clusters while addressing the noise values in a feature subset is another challenging task, which has a significant impact on final clustering. Moreover, many existing ensemble clustering techniques get input parameters, and their output is unsatisfactory for gene expression datasets.

In this paper, we propose a new algorithm to address the above-mentioned problems. Two additional steps are introduced in the clustering process of the standard fuzzy k-means algorithm. The first one is to incorporate a penalty term for making the algorithm insensitive to the initialization of cluster centroids. The second one is to automate a clustering process for iteratively updating the feature weights. This new version of fuzzy k-means algorithm is used to generate optimal number of compact feature subsets from a given high-dimensional cancer dataset. Then, it produces input clustering that contains the optimal number of clusters from each feature subsets. The weighting of features in the fuzzy k-means algorithm addresses the noise values in the dataset. We propose an ensemble clustering method, which combines a set of input clusterings into a final clustering having better overall quality. This method does not take any input parameter and calculates the relationship between clusters and shows these relationships as a tree structure. This can be helpful to the laboratory biologists for detailed analysis. Experiments on real cancer gene expression profiles illustrate that the proposed algorithm outperformed the well-known state-of-the-art tumour clustering algorithms.

The rest of this paper is organized as follows: In Section 2, we present the extended version of the fuzzy k-means clustering algorithm. Section 3 presents the proposed ensemble clustering method. In Section 4, we illustrates the experimental results. Section 5 summarizes the conclusion part.

2. The extended fuzzy k-means algorithm

Let $D = \{D_1, D_2, \ldots, D_x\}$ is a set of x instances (objects). An instance $D_i = [D_{i,1}, D_{i,1}, \ldots, D_{i,y}]$ can be expressed by different y numerical dimensions. The procedure of partitioning D into k partitions using the fuzzy k-means algorithm (Khan et al., 2019) is achieved by minimizing the following objective function:

$$P\left(C, O, V\right) = \sum_{l=1}^{k}\sum_{j=1}^{x}\sum_{i=1}^{y} o_{lj} d\left(c_{li}, d_{ji}\right) + \alpha \sum_{l=1}^{k}\sum_{j=1}^{x} o_{lj} \log o_{lj} \tag{1}$$

subject to

$$\sum_{l=1}^{k} o_{lj} = 1, \quad o_{lj} \in \left(0, 1 \quad 1 \leqslant j \leqslant x\right.$$

where

- O represents the $x \times k$ partition matrix, where each value o_{lj} shows the membership of the j_{th} instance d_j to the l_{th} partition c_l ;
- $C = \{C_1, C_2, \ldots, C_k\}$ represents the $k \times m$ matrix that contains the centroids;
- $d(z_{li}, x_{ji})$ represents the distance between the l_{th} centroid and the j_{th} instance on the i_{th} dimension.

$$d\left(c_{li}, d_{ji}\right) = \left(c_{li} - d_{ji}\right)^2 \tag{2}$$

The extension of the Fuzzy k-means algorithm using Eq. (1) is described as follows:

$$P\left(C, O, V\right) = \sum_{l=1}^{k}\sum_{j=1}^{x}\sum_{i=1}^{y} o_{lj} v_{li} d\left(c_{li}, d_{ji}\right) + \alpha \sum_{l=1}^{k}\sum_{j=1}^{x} o_{lj} \log o_{lj} + \beta \sum_{l=1}^{k} \times \sum_{i=1}^{y} v_{li} \log v_{li} \tag{3}$$

subject to

$$\sum_{i=1}^{y} v_{li} = 1, \quad v_{lj} \in \left(0, 1\right\}, \quad 1 \leqslant i \leqslant y$$

The new objective function makes the algorithm insensitive to the initial centroids and recovers the number of clusters in the data containing noise values.

Eq. (3) can be solved by addressing the following tasks:

- Task T_1: Fix $O = \widehat{O}$ and $V = \widehat{V}$, solve the reduced task $T(C, \widehat{O}, \widehat{V})$;
- Task T_2: Fix $C = \widehat{C}$ and $V = \widehat{V}$, solve the reduced task $T(\widehat{C}, O, \widehat{V})$;
- Task T_3: Fix $V = \widehat{C}$ and $O = \widehat{O}$, solve the reduced task $T(\widehat{C}, \widehat{O}, V)$;

Task T_1 is solved by

$$c_{lj} = \frac{\sum_{j=1}^{x}\sum_{i=1}^{y} o_{lj} v_{li} d_{ji}}{\sum_{j=1}^{x}\sum_{i=1}^{y} o_{lj} v_{li}} \quad \text{for} 1 \leqslant l \leqslant k \tag{4}$$

Task T_2 can be solved by using Theorem 1 as.

Theorem 1. *Let $C = \widehat{C}$ and $V = \widehat{V}$ are constant and $T(\widehat{C}, O, \widehat{V})$ can be minimized if*

$$o_{lj} = \frac{\exp\left(\frac{-P_{lj}}{\alpha}\right)}{\sum_{l=1}^{k}\exp\left(\frac{-P_{li}}{\alpha}\right)} \tag{5}$$

where

$$P_{lj} = \sum_{i=1}^{y} v_{li}\left(c_{li} - d_{ji}\right)^2 \tag{6}$$

Proof. We utilize the Lagrangian multiplier method to obtain as

$$\widetilde{X}\left(Y, \lambda\right) = \sum_{l=1}^{k}\sum_{j=1}^{x}\left(o_{lj}P_{lj} + \alpha o_{lj}\log\left(o_{lj}\right)\right) + \lambda_i\sum_{j=1}^{x}\left(\sum_{l=1}^{k} o_{lj} - 1\right) \tag{7}$$

where $\lambda = [\lambda_1, \ldots, \lambda_n]$ shows the vector that has the Lagrangian multipliers. If $(\widehat{Y}, \widehat{\lambda})$ reduces $\widetilde{X}(Y,\lambda)$, its gradients must be disappeared in both set of dimensions. Hence,

$$\frac{\Psi\widetilde{X}\left(\widehat{Y}, \widehat{\lambda}\right)}{\Psi\widehat{o}_{lj}} = P_{lj} + \alpha\left(1 + \log o_{lj}\right) + \widehat{\lambda}_i = 0 \quad \text{for} \quad 1 \leqslant l \leqslant k \text{ and } 1 \leqslant j \leqslant x \tag{8}$$

and

$$\frac{\Psi\widetilde{X}\left(\widehat{Y}, \widehat{\lambda}\right)}{\Psi\widehat{o}_{lj}} = \sum_{l=1}^{k} o_{lj} - 1 = 0 \tag{9}$$

From (8), we obtain

$$o_{lj} = \exp\left(\frac{-P_{lj}}{\alpha}\right)\exp\left(-1\right)\left(\frac{\widehat{\lambda}_i}{\alpha}\right) \tag{10}$$

We substitute (10) into (9)and obtain

$$\sum_{l=1}^{k} = \sum_{l=1}^{k}\exp\left(\frac{-P_{lj}}{\alpha}\right)\exp\left(-1\right)\exp\left(\frac{\widehat{\lambda}_i}{\alpha}\right) = \exp\left(-1\right)\exp\left(\frac{\widehat{\lambda}_i}{\alpha}\right)\sum_{l=1}^{k}\exp\left(\frac{-P_{lj}}{\alpha}\right) = 1$$

It follows as

$$o_{lj} = \frac{\exp\left(\frac{-P_{lj}}{\alpha}\right)}{\sum_{l=1}^{k}\exp\left(\frac{-P_{li}}{\alpha}\right)}$$

Task T_3 can be solved by using Theorem 2.

Theorem 2. *Let $C = \widehat{C}$ and $O = \widehat{O}$ are constant, $T(\widehat{C}, \widehat{O}, V)$ is minimized iff*

$$v_{lh} = \frac{\exp\left(\frac{-P_{lh}}{\beta}\right)}{\sum_{i=1}^{y}\exp\left(\frac{-P_{li}}{\beta}\right)} \tag{11}$$

where

$$P_{lh} = \sum_{j=1}^{x} r_{lj}\left(c_{lh} - d_{jh}\right)^2 \tag{12}$$

The proof of Theorem 2 is similar as the proof of Theorem 1.

In (3), the first part represents the k-means algorithm. The second part $\sum_{l=1}^{k} o_{lj}\log o_{lj}$ increases the membership entropy of instances to clusters. To minimize $-\sum_{l=1}^{k} o_{lj}\log o_{lj}$ is equal to maximize $\sum_{l=1}^{k} o_{lj}\log o_{lj}$. Therefore, we used $\sum_{l=1}^{k} o_{lj}\log o_{lj}$ as a penalty term in the objective function. This part is the fuzzy clustering and the membership value o_{lj} of an object is selected between zero and one. if o_{lj} value near zero for each $l \neq l^*$ and o_{jl^*} value is near one, the entropy is closer to zero. In this situation, the j_{th} instance is allocated to the l^*_{th} partition. However, the entropy becomes positive when o_{lj} values are similar for some partitions and some o_{lj} values are near zero for other set of partitions. Hence, the process of clustering can gradually minimize the intra-cluster similarity of partitions and increase the membership entropy of instances to partitions to find clusters. In this situation, the input parameter α plays a significant role. The large value of α increases the sum of negative entropies of all instances to assign each instance to more partitions and the small value of α causes to reduce the within-cluster scattering.

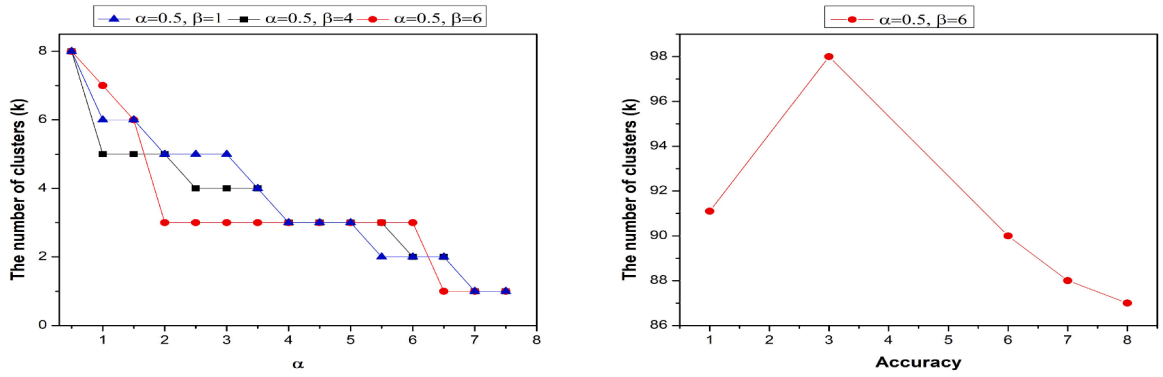
The third part in (3) introduces the feature weighted entropy that causes in the reduction of within-cluster scattering and the increase of negative weighted entropy to stimulate a large number of features for the contribution to identify clusters. The input parameter β is utilized to control the distribution of feature weights. When β is zero then the clustering process generates a result containing an important feature in every partition. When β is less than zero, o_{li} is proportional to P_{li}, the higher P_{li}, the higher o_{li}. This is inappropriate for feature weighting. Therefore, β should always be greater than zero. When β is greater then zero, o_{li} is inversely proportional to P_{li}, the smaller P_{li}, the larger o_{li}, the more important the corresponding feature.

The following section provides the detailed implementation of the extended version of the fuzzy k-means clustering algorithm to estimate the optimal number of clusters from a given noisy dataset.

2.1. The implementation

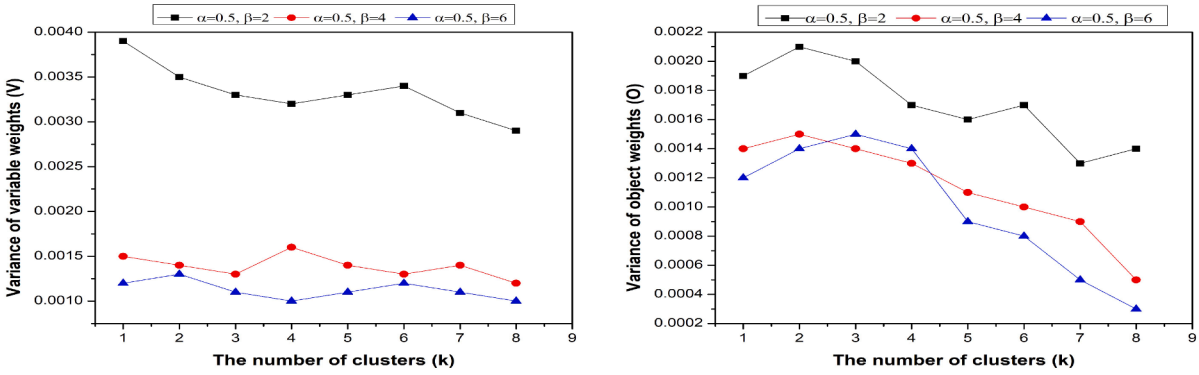
The idea of estimating the number of clusters using a given noisy dataset is summarized in Algorithm 1. The inputs of the algorithm are β, α, and k. The value of k should be greater than the actual number of cluster in the dataset. The parameter β controls the distribution of feature weights in the clustering process. The to parameters α and k are automatically learnt by the algorithm.

The algorithm comprises two loops. One loop finds the penalty factor α_{min} for generating exact k_i clusters. The other loop changes the values of k and α in decreasing and increasing orders, respectively. It can been seen that α increases from $\alpha_{min} := \alpha := \alpha_{min} \times$ h, where h = 2,3,…, and apply fuzzy k-means algorithm for each α. This loop make sure that



(a) (b)

Fig. 1. (a) The obtained results by proposed algorithm using the input parameters k = 8 and $\alpha = 0.5$ vs. different values of β. Each line represents the different numbers of merged cluster centroids obtained by changing of α and k by the algorithm against a β value. The number of merged cluster centroids provide one clustering solution. (b) The computed accuracies of clustering solutions for $\alpha = 0.5$ and $\beta = 6$.



(a) (b)

Fig. 2. The variances of object weights and feature weights against the merged cluster centres.

cluster centroids are examined, and k_s cluster centres are deleted. Therefore, the output is the number of clusters $k = k_i - k_s$. This procedure continues until the value of k becomes one. In this way, α becomes very high that all the instances belong to one merged cluster center.

This algorithm can follow different ways to determine the optimal number of clusters. If k does not change in few iterations while the values of α are increasing, this represents that the number of clusters have been estimated. The other way is to use cluster validation to verify the clustering, where a partition that shares its centroid with other partitions is identified. If the value of the number of clusters is one in the second loop, all cluster centroids reach to the similar location. The obtained output of the implementation is a set of clusterings. The clustering which gives better score of validation index provides the optimal number of clusters.

Algorithm 1. The proposed algorithm implementation.

Input: k_i (should have a maximum value, greater than the real number of clusters), $k_s = 0$ (Total clusters that share the same locations with other cluster centroids), β, $\alpha_{min} = 0.1$, iterations h = 1;
1: **while** $k_i \neq k_s$ **do**
2: Apply fuzzy k-means with (3) with k_i and β to find k_s;
3: $k_h = k_i - k_s$;
4: $\alpha_{min} = \alpha_{min}/5$;
5: **end while**
6: **while** $k_h \neq 1$ **do**
7: $\alpha = \alpha_{min} \times 1$;
8: Apply fuzzy k-means with (3) and find the number of clusters $k_h = k_i - k_s$;

(*continued on next column*)

(*continued*)

9: h = h + 1;
10: Calculate the validity measure of the current clustering and save the results (both the clustering results and the validity measure);
11: **end while**
12: **return**A clustering with the minimum validity measure.

The proposed algorithm is not sensitive the initialization of the initial cluster centroids. This gives more consistent clustering results in multiple clustering runs. The obtained results can be evaluated to determine the number of clusters. The proposed algorithm does not require the exact number of clusters in advance. This is determined by the total number of finally merged cluster centres.

2.2. *The properties of aalgorithm*

The motivation for the development of the proposed algorithm is to discover the compact feature subsets and generate input clustering that contains the optimal number of clusters from each feature subsets for ensemble clustering using the high-dimensional cancer dataset. We utilized synthetic data that contain well-structured clusters and noise dimensions to explore the properties of algorithm. We investigated the behaviors of input parameters and showed the accuracy with the appropriate estimated values of input parameters. We also presented the analysis of the two type of weights in the algorithm.

The synthetic dataset has 15 dimensions and 300 objects. The objects are categorized into three partitions with normal distribution in the first 10 dimensions. Each cluster contains 100 objects. The remaining five

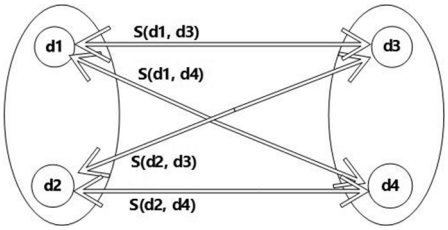


Fig. 3. Inter-cluster similarity.

features show uniformly distributed noise data. The mean values of three partitions in each feature were set as $\mu_1 = 3$, $\mu_2 = 6$, $\mu_3 = 9$, respectively, and set the constant variance $\sigma = 1$ for all partitions. These three partitions were generated separately and combined into one dataset. The noisy dimensions has uniform distribution between one and zero.

Fig. 1(a) shows the generated results by the proposed algorithm. Each line represents different clustering results generated with parameters $k = 8$, $\alpha = 0.5$, and a unique value of β. We randomly initialized the centroids as the proposed algorithm is inventive to the initiation of initial cluster centres. For the given value of β, each line shows the merging of clusters obtained by increasing the value of α and decreasing the value of k by the algorithm. When the value of α increased to certain level then the number of cluster became one. It represents the dominance of the negative entropy terms in the clustering process. If $\alpha = 2$ and $\beta = 6$, the number of clusters has not been changed in the consecutive iterations until the value of α become 6. It represents that the accurate number of clusters which is three. This can be verified by the computed accuracies of these clustering results. Fig. 1(b) shows the computed results. It can be seen that the accuracy of the algorithm is highest when the number of clusters was three.

Fig. 3 shows the analysis of weights V and O in the clustering process. As it can be seen in Fig. 2,3(a) that the variances of feature weights (V) deviate with the increase of k and α. The significant results are being produces by the given large value of β. Fig. 2(b) shows the similar effect for the variances of object weights (O). We can summarizes that the algorithm can control these weights by adjusting the input parameters α and β as follows:

- If β is large, many dimensions incorporate to the clustering; the small β causes only significant features incorporate to the clustering process.
- If α is large, each object can be assigned to more clusters by minimizing the within-cluster dispersion; the small α causes to maximize the within-cluster dispersion while the algorithm attempts to minimize it.

In the following section, we present ensemble clustering method using extended version of the fuzzy k-means algorithm for analysis of cancer data.

3. Ensemble clustering

Given a hight-dimensional cancer dataset D with a set of y features $F = \{f_1, f_2, \ldots, f_y\}$, the ensemble clustering comprises of three steps. In the first step, we get transpose of the given dataset D^t and apply Algorithm 1 to obtain e feature subsets $D^t = \{F_1^t, F_2^t, \ldots, F_e^t\}$ where $D = \cup_{i=1}^{e} F_i$. In the second step, we apply Algorithm 1 again on each feature subset F_i to obtain a set of input e input clusterings. The obtain input clusterings are more diverse in nature which increases the performance of ensemble method (Kuncheva and Hadjitodorov, 2004). In the third step, we apply Divisive Clustering Ensemble with Automatic Cluster Number (DICLENS) (Mimaroglu and Aksehirli, 2011) on input clusterings to combine them into a final clustering.

The proposed ensemble method (DICLENS) organizes all the clusters that belong to input clusterings in a minimum spanning tree. In the tree, all the vertices represent an input clusters and each edge is the inter-cluster similarity between incident vertices. This similarity is defined as follows:

$$E\left(C_{xm}, C_{yn}\right) = \frac{1}{\left|C_{xm}\right|\left|C_{yn}\right|} \sum_{a=1}^{|C_{xm}|} \sum_{b=1}^{|C_{yn}|} s\left(d_a, d_b\right) \qquad d_a \in C_{xm}, \quad d_b \in C_{yn} \tag{13}$$

In (13), $s(d_a, d_b)$ represents the number of times of that objects d_a and d_b occur in the same clusters, and $|C_{yn}|$ represents the total number of objects in cluster C_{yn} from input clustering π_y. This computation is performed using the given input clusterings which is known as evidence accumulation (Fred and Jain, 2005). For instance, the inter-cluster similarity between two clusters is shown in Fig. 3 which is computed as

$$E\left(C_{xm}, C_{yn}\right) = \frac{1}{4}\left(s(d_2, d_3) + s(d_2, d_4) + s(d_1, d_3) + s(d_1, d_4)\right) \tag{14}$$

The proposed ensemble clustering method is summarized in Algorithm 2.

Algorithm 2. The proposed ensemble clustering method.

Input: A set of input clusterings $\Pi = \{\pi_1, \pi_2, \ldots, \pi_e\}$
Output: A final clustering π_f;
1: Initialize graph $G(V = 0, E = 0, W = 0)$;
2: **for** $\pi_x \in \Pi$ $x = 1$ to e **do**
3: **for** $C_{xj} \in \pi_x$ $j = 1$ to k **do**
4: $V = V \cup C_{xj}$;
5: **end for**
6: **end for**
7: **for** $C_{xm} \in V$ $x = 1$ to e **do**
8: **for** $C_{yn} \in V$ $y = 1$ to e, $x \neq y$, $m \neq n$ **do**
9: $E = E \cup (C_{xm}, C_{yn})$;
10: $W = W \cup (E, S(C_{xm}, C_{yn}))$;
11: **end for**
12: **end for**
13: Minimum-Spanning-Tree = Prims-Algorithm (G);
14: π_f = final-clustering (Minimum-Spanning-Tree);
15: return π_f;

In Algorithm 2, all input clusterings are transformed into a weighted graph G. Each edge in the graph shows a cluster and each edge is the similarity (s) between two clusters. We apply the Prims Algorithm on graph G to get similarity-based minimum-cost spanning tree. However, there exists an inverse relationship between cost of edges and edge weight values. Since the smaller similarity shows large dissimilarity. In the following Section, we present a detailed description of final-clustering function (Line 14).

3.1. The final clustering

The inter-cluster similarity with respect to the given input clustering π_x is as follows

$$E_\pi\left(\pi_x\right) = \frac{1}{\binom{|\pi_x|}{2}} \sum_{m=1}^{|\pi_x|} \sum_{n=2}^{|\pi_x|} E\left(C_{xm}, C_{xn}\right) \qquad m \neq n \tag{15}$$

where $|\pi_x|$ denotes the total number of clusters in input clustering π_x. The large value of $E(\pi_x)$ represents that the clusters are similar, that is not required. Well-separated clusters exhibit the small value of $E(\pi)$.

The compactness of a cluster represents intra-cluster similarity, which is the similarity of pair of different objects within a cluster. It is defined as follows

Table 1
Datasets of cancer gene expression profiles.

Name	Source	n	m	# Classes
Armstrong-2002	Armstrong et al. (2002)	72	2194	3
Dyrskjot-2003	Dyrskjøt et al. (2003)	40	1203	3
Risinger-2003	Risinger et al. (2003)	42	1771	4
Yeoh-2002	Yeoh et al. (2002)	248	2526	6
Bhattacharjee-2001	Bhattacharjee et al. (2001)	203	1543	5
Chowdary-2006	Chowdary et al. (2006)	104	182	2
Gordon-2002	Gordon et al. (2002)	181	1626	2
Laiho-2007	Laiho et al. (2007)	37	2202	2

$$I\left(C_{xn}\right) = \frac{1}{\binom{|C_{xn}|}{2}} \sum_{a=1}^{|C_{xn}|} \sum_{b=2}^{|C_{xn}|} s\left(d_a, d_b\right) \quad a \neq b \tag{16}$$

where $|C_{xn}|$ is the total number of objects in cluster C_{xn} and $s(d_a, d_b)$ is the number of times that objects d_a and d_b appears to the same clusters. The large value of $I(C_{xn})$ represents the more compact cluster.

The intra-cluster similarity of an input clustering π_x is defined as follows

$$I_\pi\left(\pi_x\right) = \frac{1}{|\pi_x|} \sum_{n=1}^{|\pi_x|} I\left(C_{xn}\right) \quad C_{xn} \in \pi_x \tag{17}$$

where $|\pi_x|$ represents the total number of clusters in input clustering π_x. The aim is to generate more compact and well-separated clusters in a final clustering using the given input clusterings. We can determine the quality of an input clustering π_x as follows

$$\Phi\left(\pi_x\right) = \overline{I_\pi}\left(\pi_x\right) - \overline{E_\pi}\left(\pi_x\right) \tag{18}$$

where $\overline{I_\pi}(\pi_x)$ and $\overline{E_\pi}(\pi_x)$ are the normalized values of $I_\pi(\pi_x)$ and $E_\pi(\pi_x)$, respectively. This normalization controls the effect of the large values of both measurements. The maximum values always dominant in the results while the small ones are unable to effect the outcome. These values are computed very fast—linear using the total number of clusters in the input clusterings as stated in Khan et al. (2016).

The final clustering π_f is the one which maximizes the quality function as

$$\pi_f = \underset{\pi_x \in \Pi}{\operatorname{argmax}} \Phi\left(\pi_x\right) \tag{19}$$

The procedure of obtaining final clustering is summarized in Algorithm 3.

Algorithm 3. Final-clustering (Minimum-Spanning-Tree)

Input: Similarity based Minimum Spanning Tree (*MST*);
Output: A final clustering;
1: Initialize $I_\pi(\pi_k)$, $E_\pi(\pi_k)$, $\overline{I_\pi}(\pi_k)$, $\overline{E_\pi}(\pi_k)$, Φ;
2: **for** k == 1 to *No. of edges in MST* **do**
3: Delete an edge with minimum weight from *MST*;
4: π_m is a meta-clustering and each connected component is a meta-cluster;
5: π_k = majority-voting(π_m);
6: Compute $I_\pi(\pi_k)$;
7: Compute $E_\pi(\pi_k)$;
8: **end for**
9: $\overline{I_\pi}(\pi_k)$ = normalize ($I_\pi(\pi_k)$);
10: $\overline{E_\pi}(\pi_k)$ = normalize($E_\pi(\pi_k)$);
11: **for** i == 1 to k
12: $\Phi_i = \overline{I_\pi}(\pi_k) - \overline{E_\pi}(\pi_k)$;
13: **end for**
14: *MIndex* = $Max_i(\Phi)$;
15: return π_{MIndex};

Algorithm 4. Normalize (list)

Input: A list;
Output: A normalized list;
1: max = max(*list*); min = min(*list*); *normalized* − *list* = 0;
2: **for** i = *Start*(*list*) to *End*(*list*) **do**
3: $normalized\ list = \frac{list_i - min}{max - min}$;
4: **end for**
5: return *normalized list*;

Algorithm 3 shows the procedure of finding the robust final clustering π_f, that contains more compactness and separateness of clusters. In this procedure, we start from the minimum edge weight and cut an edge in MST agglomeratively and generate a meta-clustering. Each meta-cluster is an attached part of the tree. In the tree, the edges with minimum weights are good candidate for deletion because they represent weak linkages. After every iteration, the connected clusters seems more similar which shows that the objects in meta-cluster coexist frequently in input clusterings. However, each meta-cluster has a set of clusters, an object can exist many times in a meta-cluster. Therefore, we performed majority voting on the meta-clustering to avoid overlapping in final clustering. Consequently, an object is assigned to that cluster where that exist many times.

4. Experimental results

In this section, we present a series of experiments on eight real-world cancer gene expression profiles to evaluate the proposed algorithm and other cluster ensembles techniques.

4.1. Real-world datasets

Table 1 shows the details of gene expression profiles, where n is the number of records and m is the number of features in a dataset. The Armstrong-2002 data is categorized into lymphoblastic leukaemias with MLL translocations (MLL), conventional acute lymphoblastic (ALL) and acute myelogenous leukaemia (AML) (Yu et al., 2014). The Dyrskjot-2003 dataset is utilized to find the clinically relevant subclasses of bladder carcinoma. The Risinger-2003 dataset is utilized to find normal endometria samples from three categories of cancer samples. These categories are serous papillary non-endometrioid cancer samples, clear cell non-endometrioid cancer samples, and endometrioid cancer samples. The Yeoh-2002 dataset has 248 pediatric acute lymphoblastic leukemia (ALL) samples, which can be grouped into six various subtypes (clusters). There are 203 samples in the Bhattacharjee-2001 dataset which can be used to separate the normal lung tissues (NL) from four types of lung tumour samples. These samples are adenocarcinoma (AD), pulmonary carcinoids (COID), squamous cell lung carcinomas (SQ), and small cell lung cancer (SCLC) respectively. The Chowdary-2006 dataset contains lymph node-negative breast tumours samples and Dukes B colon tumours samples. The samples of Gordon-2002 dataset can be categorized into malignant pleural mesothelioma (MPM) and adenocarcinoma (AD) of the lung. The Laiho-2007 dataset contains serrated colorectal carcinomas (CRCs) and conventional CRCs samples. We used the same preprocessing of the datasets as that in de Souto et al. (2008).

4.2. Experimental settings and evaluation measures

The performances of the cluster ensemble approaches are measured by the rand index (RI) and the purity (PU) measure. These evaluation measure are most commonly used for cluster analysis. The rand index (Rand, 1971) is defined as

$$RI = \frac{2(t_1 + t_4)}{(t_1 + t_2 + t_3 + t_4)(t_1 + t_2 + t_3 + t_4 - 1)} \tag{20}$$

Table 2
Comparisons of the generated results by the Proposed Algorithm (PA), traditional clustering algorithms, and the fuzzy extension models in RDCFCE with respect to PU. The measurements are mean values and maximum values of the obtained results by running each algorithm ten times.

Dataset	M_1	M_2	M_3	M_4	PA	KM	HCl
Armstrong-2002	0.90±0.93	0.89±0.91	0.92±0.96	0.93±0.97*	0.93±0.96	0.86±0.89	0.83±0.87
Dyrskjot-2003	0.72±0.75	0.71±0.72	0.66±0.73	0.76±0.81	0.83±0.85*	0.67±0.70	0.65±0.68
Risinger-2003	0.76±0.79	0.74±0.78	0.63±0.72	0.76±0.80	0.79±0.83*	0.61±0.67	0.58±0.61
Yeoh-2002	0.57±0.60	0.57±0.58	0.56±0.63	0.58±0.68	0.69±0.73*	0.47±0.51	0.52±0.54
Bhattacharjee-2001	0.89±0.91	0.84±0.86	0.89±0.92*	0.83±0.86	0.88±0.91	0.74±0.77	0.68±0.70
Chowdary-2006	0.94±0.96	0.62±0.66	0.95±0.97	0.96±0.97	0.98±0.98*	0.71±0.74	0.66±0.69
Gordon-2002	0.97±0.98	0.91±0.96	0.94±0.96	0.97±0.98*	0.97±0.98*	0.88±0.91	0.87±0.98
Laiho-2007	0.81±0.87	0.83±0.88	0.79±0.81	0.88±0.90	0.92±0.94*	0.68±0.73	0.71±0.74

Table 3
Comparisons of the generated results by the Proposed Algorithm (PA), traditional clustering algorithms, and the fuzzy extension models in RDCFCE with respect to RI. The measurements are mean values and maximum values of the obtained results by running each algorithm ten times.

Dataset	M_1	M_2	M_3	M_4	PA	KM	HCl
Armstrong-2002	0.89±0.92	0.87±0.88	0.92±0.93	0.95±0.97	0.96±0.97*	0.88±0.91	0.86±0.89
Dyrskjot-2003	0.74±0.81	0.72±0.78	0.65±0.69	0.71±0.80	0.82±0.86*	0.72±0.75	0.69±0.73
Risinger-2003	0.73±0.75	0.72±0.75	0.66±0.71	0.72±0.81	0.80±0.83*	0.71±0.74	0.72±0.74
Yeoh-2002	0.71±0.78	0.70±0.75	0.75±0.78	0.75±0.79	0.82±0.82*	0.74±0.77	0.68±0.70
Bhattacharjee-2001	0.67±0.74*	0.64±0.67	0.67±0.71	0.65±0.67	0.72±0.73	0.63±0.67	0.64±0.66
Chowdary-2006	0.90±0.91	0.51±0.55	0.91±0.92	0.92±0.93	0.95±0.95*	0.88±0.91	0.87±0.89
Gordon-2002	0.93±0.96	0.71±0.92	0.91±0.93	0.92±0.95	0.97±0.98*	0.65±0.72	0.68±0.73
Laiho-2007	0.68±0.76	0.70±0.79	0.52±0.61	0.77±0.79	0.82±0.85*	0.61±0.68	0.66±0.79

$$t_1 = \sum_{i,h} 1\left\{\left(b_i, b_h\right) | b_i \in C_k, b_h \in C_k, b_i \in \overline{C_j}, b_h \in \overline{C_k}\right\},$$
$$t_2 = \sum_{i,h} 1\left\{\left(b_i, b_h\right) | b_i \in C_k, b_h \in C_k, b_i \in \overline{C_j}, b_h \notin \overline{C_k}\right\},$$
$$t_3 = \sum_{i,h} 1\left\{\left(b_i, b_h\right) | b_i \in C_k, b_h \notin C_k, b_i \in \overline{C_j}, b_h \in \overline{C_k}\right\},$$
$$t_4 = \sum_{i,h} 1\left\{\left(b_i, b_h\right) | b_i \in C_k, b_h \notin C_k, b_i \in \overline{C_j}, b_h \notin \overline{C_k}\right\}.$$

where $\{b_i \in C_k, b_h \in C_k, b_i \in \overline{C_j}, b_h \in \overline{C_k}\}$ represents that the samples b_i and b_h belong to cluster C_k of real partition are both assigned to cluster $\overline{C_k}$ of predicted partition. The higher the value of RI indicates the batter clustering result.

The The purity measure is defined as

$$PU\left(C, \overline{C}\right) = \frac{1}{n}\sum_{i=1}^{k} \max_{h \in \{1,\ldots,\overline{k}\}} \left(C_i \cap \overline{C_h}\right) \tag{21}$$

The large value of PU indicates the better clustering result.

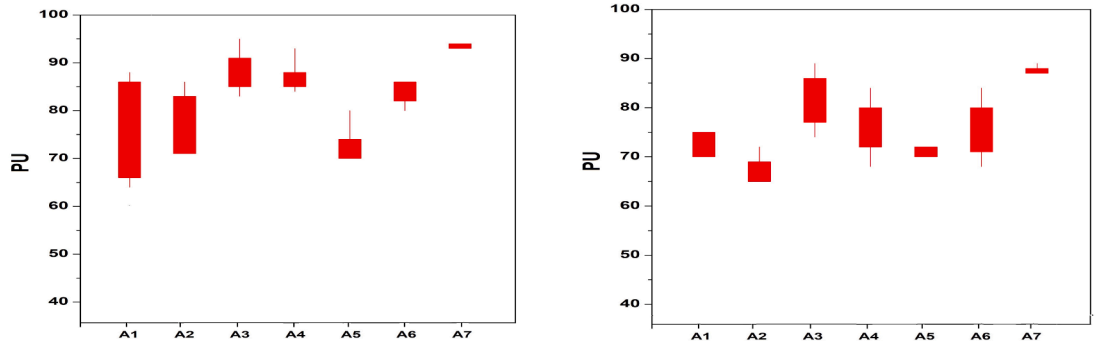
The performances of the proposed algorithm and other cluster ensemble algorithms are measured by the mean value and the maximum values of PU and RI after applying each algorithm ten times. The best results are marked with an asterisk symbol.

In the following section, we first compare the performance of proposed algorithm with the performances of different models of Random Double Clustering based Fuzzy Cluster Ensemble (RDCFCE) (Yu et al., 2014) framework. Next, we investigate the proposed algorithm and other consensus clustering approaches. Then, we present the execution time of the proposed algorithm.

4.3. Comparisons of proposed algorithm vs. RDCFCE

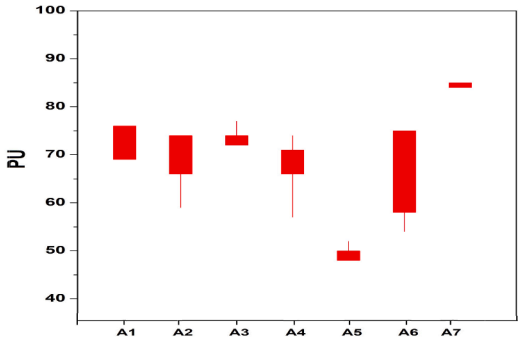
In this section, we use cancer data to explore the performance of the proposed algorithm. We compare the obtained results with that of RDCFCE based models. These models are derived by adopting fuzzy extension model in RDCFCE. They have two-value based linear extension model (M1), the two-value based nonlinear extension model (M2), the three-value based extension model (M3), and the newly proposed representative extension model (M4) (Yu et al., 2014). We also used two well-known traditional clustering algorithms for the comparisons. The are the K-Means algorithm (Likas et al., 2003) and the Hierarchical (single linkage) Clustering (HCl) algorithm (Johnson, 1967).

Table 2, 3 show the comparison results of the proposed algorithm and RDCFCE based models with respect to PU and RI on the different datasets. The proposed algorithm achieved significant performance on the majority of the eight cancer gene datasets. These profiles are the Dyrskjot-2003 dataset, the Risinger-2003 dataset, the Yeoh-2002 dataset, the Chowdary-2006 dataset, Gordon-2002, and the Laiho-2007 dataset. The best result of the proposed algorithm is obtained on the Chowdary-2006 dataset under PU evaluation measure, and successfully assigned 104 samples to corresponding clusters. The performances of the model M_4 and that of the proposed algorithm are almost similar on Gordon-2002 dataset, both correctly divided 181 samples into their correct clusters. The model M_1 outperformed all the algorithms on Bhattacharjee-2001 dataset under RI measure, and successfully used 203 samples used to separate the normal lung tissues (NL) from four types of lung tumour samples. On the other hand, the performance of the model M_4 is also better than that of all the algorithms on Armstrong-2002 dataset under PU measure and successfully categorized 72 samples into lymphoblastic leukaemias with MLL translocations (MLL), conventional acute lymphoblastic (ALL), and acute myelogenous leukaemias (AML). A leading result is achieved by the proposed algorithm on Laiho-2007 dataset under RI evaluation measure, and correctly distinguished serrated colorectal carcinomas from conventional CRCs. The proposed algorithm also outperformed the well-known clustering algorithms on different datasets under different evaluation measures. In summary, the proposed algorithm is able to achieve good clustering performance.

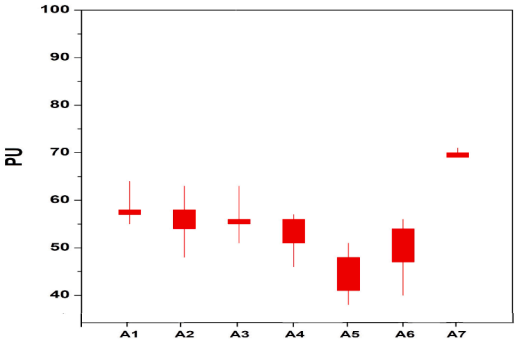


(a) The Armstrong-2002 dataset.

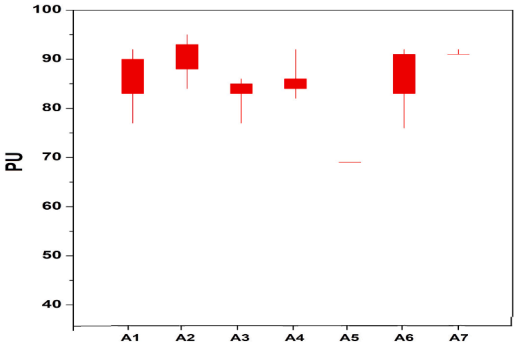
(b) The Dyrskjot-2003 dataset.



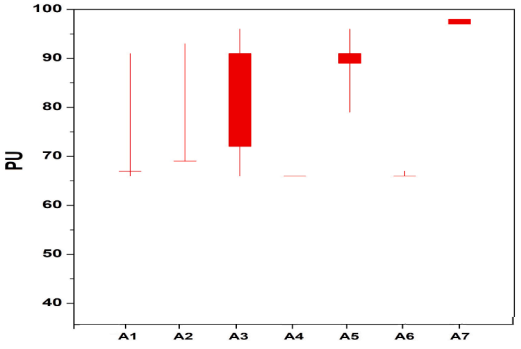
(c) The Risinger-2003 dataset.



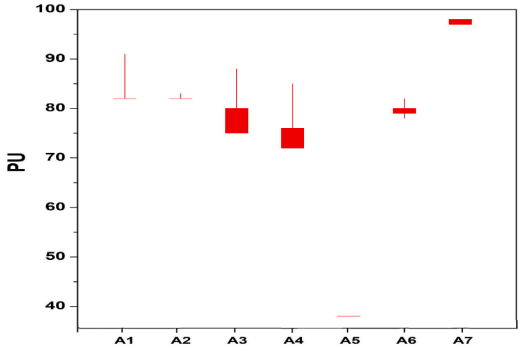
(d) The Yeoh-2002 dataset.



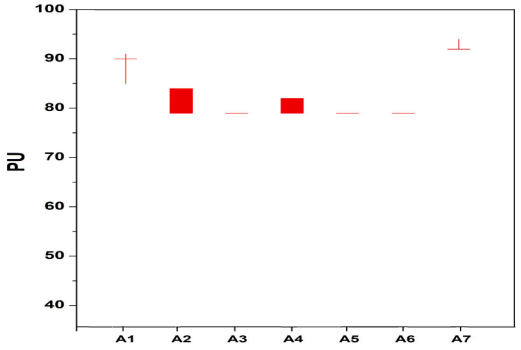
(e) The Bhattacharjee-2001 dataset.



(f) The Chowdary-2006 dataset.



(g) The Gordon-2002 dataset.

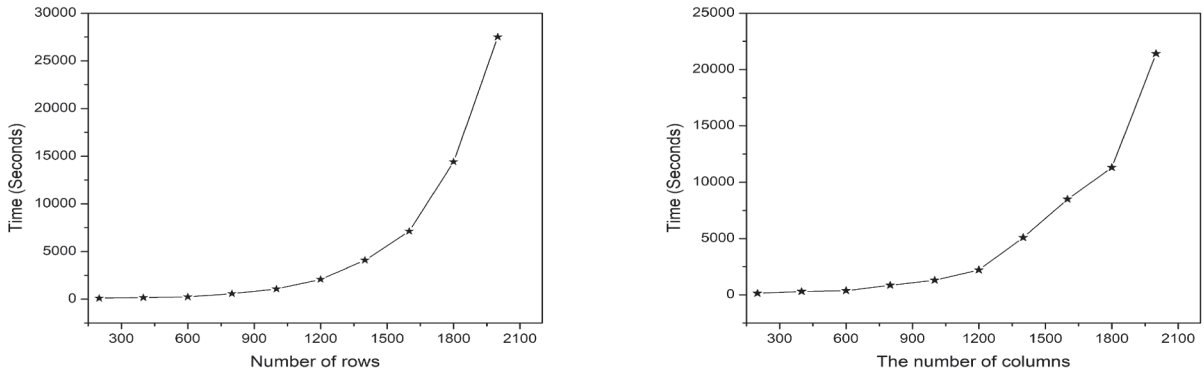


(h) The Laiho-2007 dataset.

Fig. 4. The comparisons of results by the proposed algorithm with the the obtained results by other ensemble clustering approaches under PU evaluation measure (where A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8 represent SOM^2CE, NG^2CE, RSSOM, RSNG, RSFCM, RSKM, and the proposed algorithm, respectively).

Table 4
The t-test results between the proposed algorithm and SOM^{2}CE, NG^{2}CE, RSSOM, RSNG, RSFCM, and RSKM on different datasets with respect to PU measure.

Name	SOM^{2}CE	NG^{2}CE	RSSOM	RSNG	RSFCM	RSKM
Armstrong-2002	1	1	1	1	1	1
Dyrskjot-2003	1	1	0	0	1	1
Risinger-2003	0	0	0	1	1	1
Yeoh-2002	0	0	0	1	1	1
Bhattacharjee-2001	0	1	0	0	1	0
Chowdary-2006	1	1	1	1	1	1
Gordon-2002	1	1	1	1	1	1
Laiho-2007	0	1	1	1	1	1



(a) Execution time versus the number of rows. (b) Execution time versus the number of columns.

Fig. 5. Execution time of the proposed algorithm on synthetic data.

4.4. Comparisons of proposed algorithm vs. other cluster ensemble approaches

In this section, we compare the performance of the proposed algorithm with the ones of the other well-known ensemble clustering techniques with respect to PU evaluation measure. These techniques are SOM^{2}CE (Yu et al., 2012), the double Natural Gas based cluster ensemble (NG^2) CE technique (Chen et al., 2012), the ensemble cluster technique based on the Random Sub-Space approach and the Self-Organizing Map algorithm (RSSOM) (Kohonen, 1996), the ensemble clustering technique based on the Random Sub-Space technique and the Fuzzy C-Means algorithm (RSFCM) (Yu et al., 2013), and the ensemble clustering technique based on the Random Sub-Space approach and K-Means (RSKM) (Yu et al., 2007).

Fig. 4 shows the comparisons of obtained results using the proposed algorithm and other well-known ensemble clustering approaches (where A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8 represent SOM^{2}CE, NG^{2}CE, RSSOM, RSNG, RSFCM, RSKM, and the proposed algorithm, respectively). It can be seen that the proposed algorithm achieved better performance on seven out of eight cancer datasets. This shows that the proposed algorithm is able to improve the performance on tumour clustering. The reasons are considering a significant number of input clusterings by generating an optimal number of feature subsets, generating diverse input clusterings with an optional number of clusters in each input clustering, and using an appropriate ensemble method to combine input clusterings into a final clustering in the proposed algorithm. Moreover, incorporating a step to address missing or sparse values in the data.

It can be seen in the Fig. 4, SOM^{2}CE and NG^{2}CE obtained satisfactory results on many datasets. For example, NG^{2}CE achieved the best performance on the Bhattacharjee-2001 dataset. The performance of SOM^{2}CE is good on the Laiho-2007 dataset. Both SOM^{2}CE and NG^{2}CE perform clustering analysis on both the sample and feature dimensions, which make them effective to noisy genes and samples. The obtained results from some datasets by RKSOM and RSKM are also good, since the random subspace approach is able to suppress the effect of noise in the data. In general, the proposed algorithm is a suitable choice for performing cluster analysis of gene expression profiles.

To investigate the difference between the proposed algorithm and the other ensemble clustering techniques, we used the t-test with a confidence level of 95%. Table 4 shows the computed results of t-test between the prosed algorithm and SOM^{2}CE, NG^{2}CE, RSSOM, RSNG, RSFCM, RSKM on all cancer datasets under PU evaluation measure. The results shows that the differences between the proposed algorithm and other ensemble clustering techniques are significants, particularly on Armstrong-2002, Gordon-2002, and Chowdary-2006 data. The t-test is utilized to find if the obtained results by the proposed algorithm are very different from the ones generated by other cluster ensemble techniques. The results of t-test are inevitably affected by the type I error, that is relevant to the number of tests. Since, we conducted a small number of tests in our experiments, therefore, the type I error is small.

4.5. Execution time of the proposed algorithm

To investigate the execution time of the proposed algorithm, we generated two categories of synthetic data. The first category contains the fixed number of columns 1000. We generated 10 different versions of sub-datasets by varying the number of rows in the range from 200 to 2000. We clustered each sub-dataset into three clusters with normal distribution in the first 950 features. Each cluster contains approximately the equal number of objects. The remaining 50 features show uniformly distributed noise data. The mean values of three partitions in each key feature were set as $\mu_1 = 3$, $\mu_2 = 6$, $\mu_3 = 9$, respectively, and the same variance $\sigma = 1$ was used for all partitions. These three partitions were generated independently and combined into each sub dataset, accordingly. The noisy features has uniform distribution between one and zero. We used the same procedure to create the data of second category while keeping the fixed number of rows 1000 and varying the number of columns from 10 times in the range from 200 to 2000.

Fig. 5 shows the execution time of the proposed algorithm in synthetic data. For each dataset, we set the input parameters $\alpha = 0.5$, $\beta =$

4, and $k = 10$. For each sub dataset, we used the proposed algorithm with random initialization of different cluster centroids to produce 10 input clusterings. Then, the proposed ensemble clustering method is used to a combine the input clusterings into one clustering solution. As it can be seen in the figure, the execution time of the proposed algorithm increases linearly as the numbers of rows and features increased. These results show that the proposed algorithm is scalable to large datasets.

5. Conclusions

In this paper, we presented a new ensemble clustering algorithm to cluster real cancer gene expression profiles. This algorithm extended the well-known fuzzy k-means algorithm by introducing two new steps into its objective function. The first one is to incorporate a penalty term for making the algorithm insensitive to the initialization of cluster centroids. The second one is to automate a clustering process for iteratively updating the feature weights. The feature weighting part addresses the noise values in the dataset. These steps are very useful for generating optimal number of feature subsets using a given dataset, insensitiveness of initialization of initial cluster centroids, and generating the input clustering from each feature subset by determining the optimal number of clusters. In this way, we generated divers input clusterings using the feature sub datasets of the given data. We proposed a new ensemble clustering method that combined a set of input clusterings into a final clustering having better overall quality. Experiments on real cancer gene expression profiles illustrate that the proposed algorithm works well on these data sets, and outperform most of the state-of-the-art clustering algorithms.

In our future work, we will update our algorithm that can automatically perform cluster analysis on the categorical, financial, and streaming data. Finally, we will improve and test our method on further real applications.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is partially supported by BNU startup research fund and UIC startup research fund (R72021110).

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature, 403*(6769), 503–511.

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R., & Korsmeyer, S. J. (2002). Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics, 30*(1), 41–47.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., et al. (2001). Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences, 98*(24), 13790–13795.

Chen, H., Yu, Z., Han, G., You, J. & Li, L. (2012). Ng 2 ce: Double neural gas based cluster ensemble framework. In 2012 7th International conference on computer science & education (ICCSE) (pp. 26–31). IEEE.

Chowdary, D., Lathrop, J., Skelton, J., Curtin, K., Briggs, T., Zhang, Y., Yu, J., Wang, Y., & Mazumder, A. (2006). Prognostic gene expression signatures can be measured in tissues collected in rnalater preservative. *The Journal of Molecular Diagnostics, 8*(1), 31–39.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2*, 224–227.

de Souto, M. C., Costa, I. G., de Araujo, D. S., Ludermir, T. B., & Schliep, A. (2008). Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics, 9*(1), 497.

Dyrskjøt, L., Thykjaer, T., Kruhøffer, M., Jensen, J. L., Marcussen, N., Hamilton-Dutoit, S., Wolf, H., & Ørntoft, T. F. (2003). Identifying distinct classes of bladder carcinoma using microarrays. *Nature Genetics, 33*(1), 90–96.

Fred, A. L., & Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(6), 835–850.

Fred, A. L., & Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(6), 835–850.

Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J., & Bueno, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research, 62*(17), 4963–4967.

Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics, 21*(15), 3201–3212.

Iam-On, N., Boongoen, T., & Garrett, S. (2010). Lce: A link-based cluster ensemble method for improved gene expression data analysis. *Bioinformatics, 26*(12), 1513–1519.

Iam-On, N., Boongoen, T., & Garrett, S. (2010). Lce: A link-based cluster ensemble method for improved gene expression data analysis. *Bioinformatics, 26*(12), 1513–1519.

Jaskowiak, P. A., Campello, R. J., & Costa, I. G. (2013). Proximity measures for clustering gene expression microarray data: A validation methodology and a comparative analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10*(4), 845–857.

Jing, L., Tian, K., & Huang, J. Z. (2015). Stratified feature sampling method for ensemble clustering of high dimensional data. *Pattern Recognition, 48*(11), 3688–3702.

Jing, L., Tian, K., & Huang, J. Z. (2015). Stratified feature sampling method for ensemble clustering of high dimensional data. *Pattern Recognition, 48*(11), 3688–3702.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*(3), 241–254.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*, (Vol. 344). John Wiley & Sons.

Khan, I., Huang, J. Z., & Ivanov, K. (2016). Incremental density-based ensemble clustering over evolving data streams. *Neurocomputing, 191*, 34–43.

Khan, I., Huang, J. Z., Luo, Z., & Masud, M. (2018). Cplp: An algorithm for tracking the changes of power consumption patterns in load profile data over time. *Information Sciences, 429*, 332–348.

Khan, I., Huang, J. Z., Tung, N. T., & Williams, G. (2014). Ensemble clustering of high dimensional data with fastmap projection. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 483–493.

Khan, I., Luo, Z., Huang, J. Z., & Shahzad, W. (2019). Variable weighting in fuzzy k-means clustering to determine the number of clusters. *IEEE Transactions on Knowledge and Data Engineering*.

Kim, E.-Y., Kim, S.-Y., Ashlock, D., & Nam, D. (2009). Multi-k: Accurate classification of microarray subtypes using ensemble k-means clustering. *BMC Bioinformatics, 10*(1), 260.

Kohonen, T. (1996). Emergence of invariant-feature detectors in the adaptive-subspace self-organizing map. *Biological Cybernetics, 75*(4), 281–291.

Kuncheva, L. I. & Hadjitodorov, S. T. (2004). Using diversity in cluster ensembles. In 2004 IEEE international conference on systems, man and cybernetics (IEEE Cat. No. 04CH37583) (Vol. 2, pp. 1214–1219). IEEE.

Laiho, P., Kokko, A., Vanharanta, S., Salovaara, R., Sammalkorpi, H., Järvinen, H., Mecklin, J.-P., Karttunen, T. J., Tuppurainen, K., Davalos, V., et al. (2007). Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. *Oncogene, 26*(2), 312–320.

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern Recognition, 36*(2), 451–461.

Maji, P., & Paul, S. (2012). Rough-fuzzy clustering for grouping functionally similar genes from microarray data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10*(2), 286–299.

Mimaroglu, S., & Aksehirli, E. (2011). Diclens: Divisive clustering ensemble with automatic cluster number. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9*(2), 408–420.

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning, 52*(1–2), 91–118.

Panwong, P., Boongoen, T., & Iam-On, N. (2020). Improving consensus clustering with noise-induced ensemble generation. *Expert Systems with Applications, 146*, Article 113138.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*(336), 846–850.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*(336), 846–850.

Risinger, J. I., Maxwell, G. L., Chandramouli, G., Jazaeri, A., Aprelikova, O., Patterson, T., Berchuck, A., & Barrett, J. C. (2003). Microarray analysis reveals distinct gene expression profiles among different histologic types of endometrial cancer. *Cancer Research, 63*(1), 6–11.

Sarle, W. S. (1990). Algorithms for clustering data.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

Shannon, C. E. (1951). Prediction and entropy of printed english. *Bell System Technical Journal, 30*(1), 50–64.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research, 3* (Dec):583–617.

Tautenhain, C. P., & Nascimento, M. C. (2020). An ensemble based on a bi-objective evolutionary spectral algorithm for graph clustering. *Expert Systems with Applications, 141*, Article 112911.

Yang, F., Li, X., Li, Q., & Li, T. (2014). Exploring the diversity in cluster ensemble generation: Random sampling and random projection. *Expert Systems with Applications, 41*(10), 4844–4866.

Yeoh, E.-J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., et al. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell, 1*(2), 133–143.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., & Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics, 17*(10), 977–987.

Yu, Z., Chen, H., You, J., Han, G., & Li, L. (2013). Hybrid fuzzy cluster ensemble framework for tumor clustering from biomolecular data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10*(3), 657–670.

Yu, Z., Chen, H., You, J., Li, L. & Han, G. (2012). Som 2 ce: Double self-organizing map based cluster ensemble framework and its application in cancer gene expression profiles. In International conference on industrial, engineering and other applications of applied intelligent systems (pp. 351–360). Springer.

Yu, Z., Chen, H., You, J., Liu, J., Wong, H.-S., Han, G., & Li, L. (2014). Adaptive fuzzy consensus clustering framework for clustering analysis of cancer data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 12*(4), 887–901.

Yu, Z., & Wong, H.-S. (2009). Class discovery from gene expression data based on perturbation and cluster ensemble. *IEEE Transactions on NanoBioscience, 8*(2), 147–160.

Yu, Z., Wong, H.-S., & Wang, H. (2007). Graph-based consensus clustering for class discovery from gene expression data. *Bioinformatics, 23*(21), 2888–2896.

Yu, Z., Wong, H.-S., & Wang, H. (2007). Graph-based consensus clustering for class discovery from gene expression data. *Bioinformatics, 23*(21), 2888–2896.

Zheng, C.-H., Huang, D.-S., Zhang, L., & Kong, X.-Z. (2009). Tumor clustering using nonnegative matrix factorization with gene selection. *IEEE Transactions on Information Technology in Biomedicine, 13*(4), 599–607.